Penalized Competing Risks Analysis using Case-base sampling

Nirupama Tamvada

June 21

Department of Statistics, UBC

Overview

1. Background and Motivation

2. Proposed Method

3. Simulation Study

1. Variable Selection
2. CIF Prediction

4. Discussion and Future Work

Background

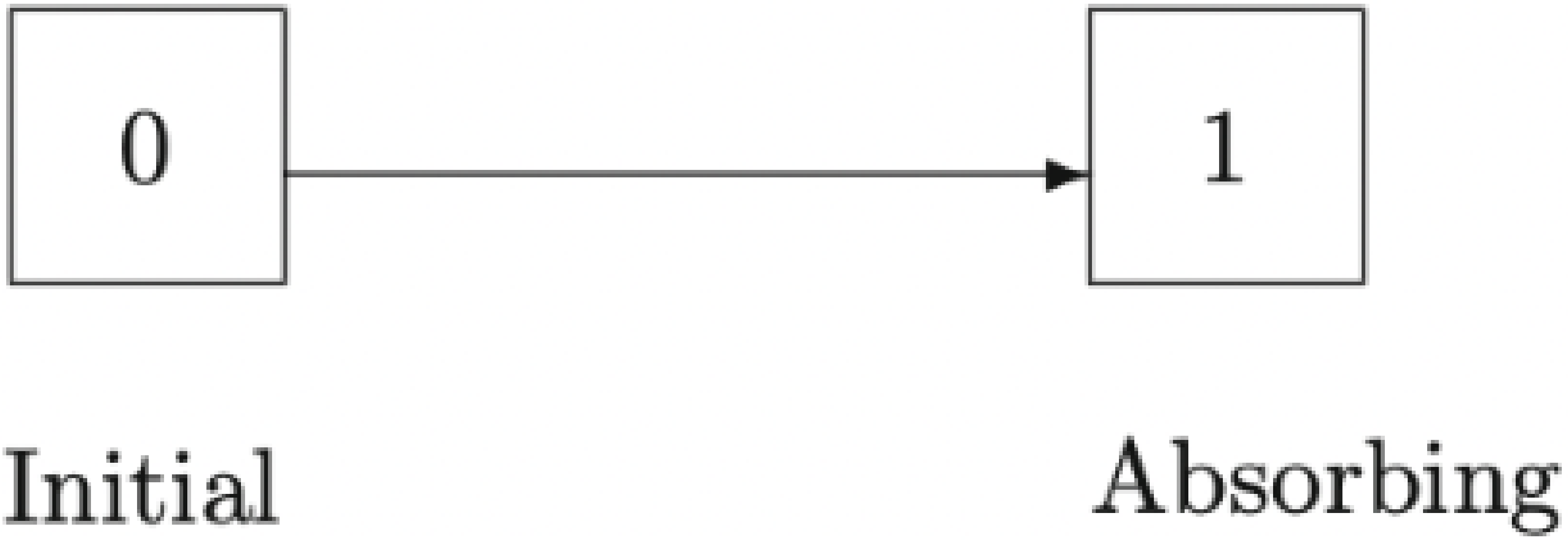
Survival Analysis

- Quantify expected time to event, e.g. death, onset of disease

Proportion of a population which will survive past a certain time?

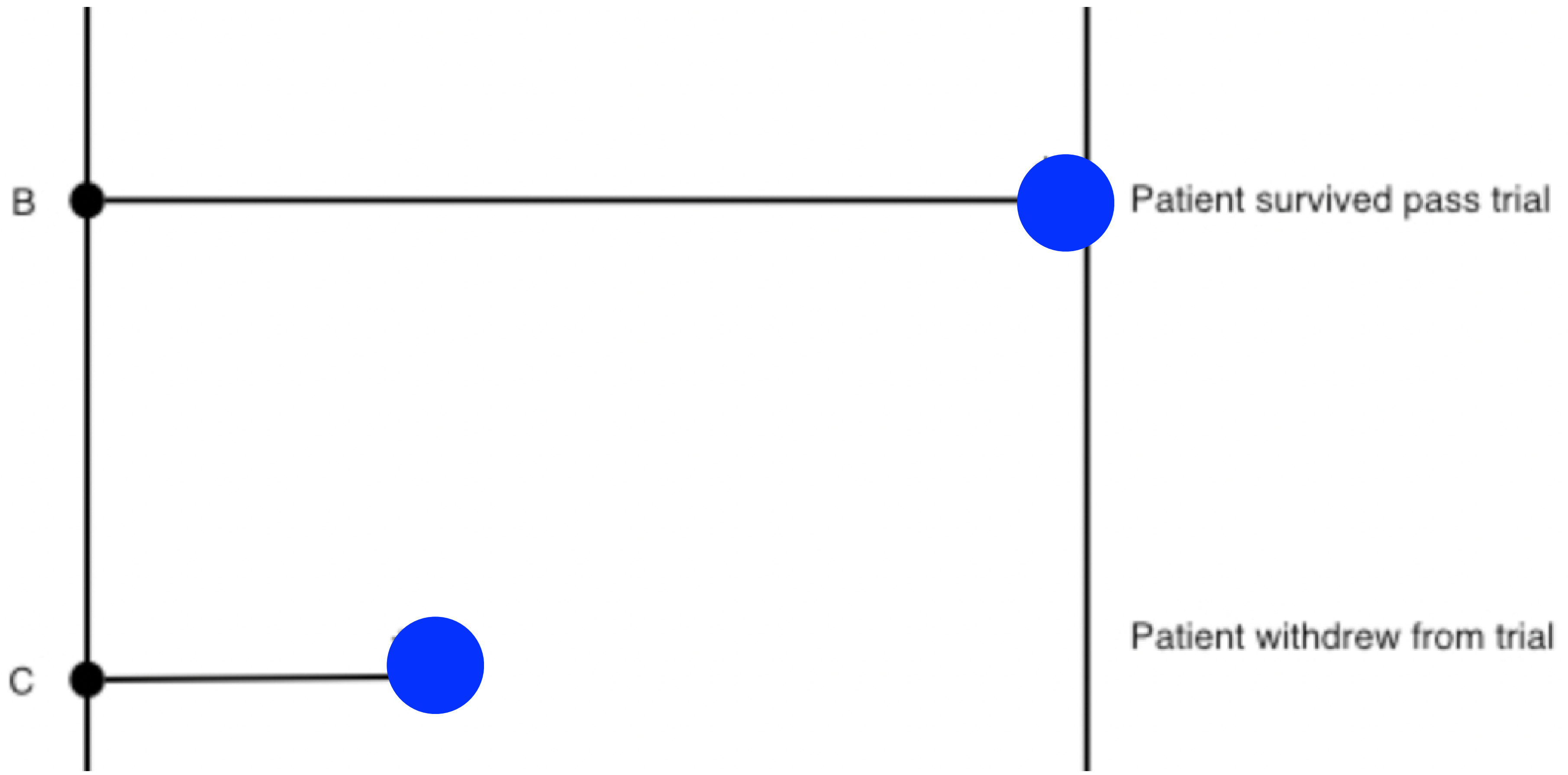
Of the surviving population, at what rate will they eventually experience the event?

How do particular characteristics affect event rate?



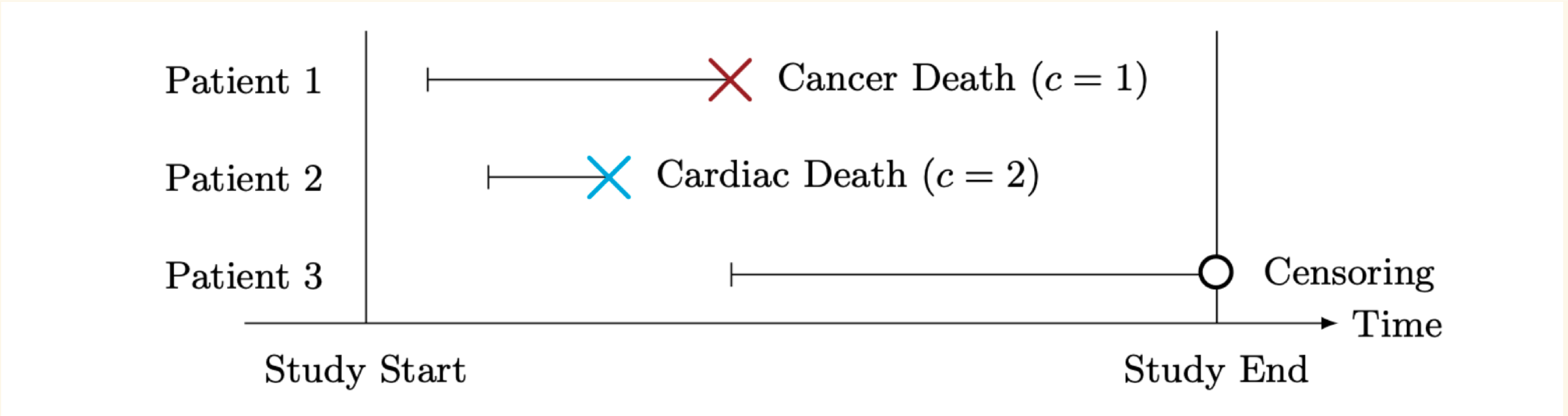
Censoring is common issue in survival analysis

- **Follow individuals typically for a fixed study period**
- When study ends, some individuals still have not experienced event yet
- Individuals drop out: event status unknown



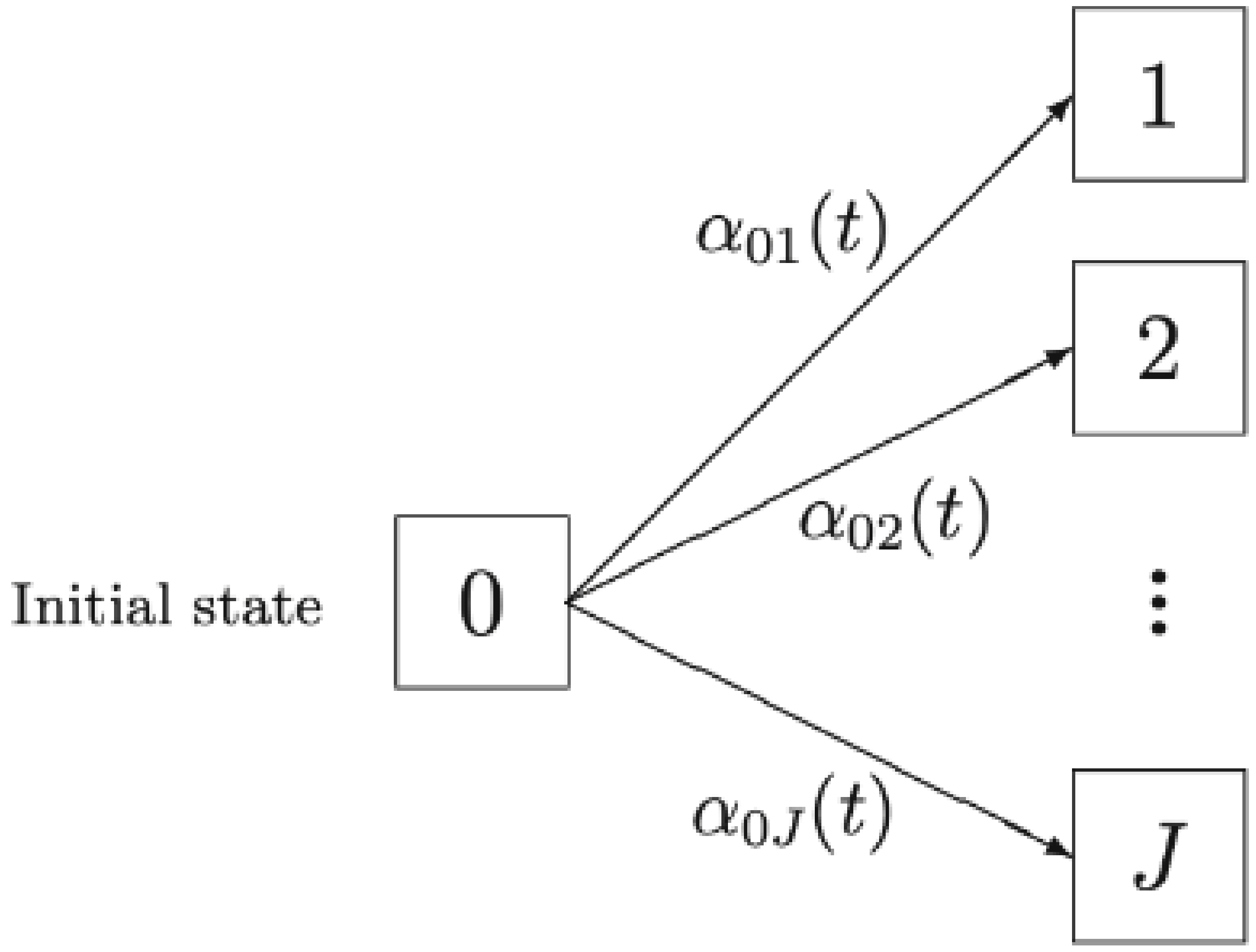
Competing Risks: Extension of Survival Analysis

- In real life, subjects can potentially experience more than one type of a certain event



Competing Risks: Extension of Survival Analysis

- Naive survival analysis: treat other event as censored
- Individual is no longer *at risk* of experiencing event of interest



Outcomes of Interest in Survival Analysis

1. Hazard

2. Cumulative incidence

Survival Analysis: Hazard

- **Describes rate:** instantaneous risk of event *J* for subjects that are currently **event-free**
- **Quantify risk factors:** *"1-unit change in Age can increase the rate of occurrence of event by 2.11"*

$$\alpha(t) \cdot dt := \lim_{\Delta t \searrow 0} \frac{P(T \in [t, t + \Delta t) | T \geq t)}{\Delta t}$$

Survival Analysis: Cumulative Incidence (CIF)

- **Describes risk:** Probability of occurrence of event of interest/*incidence* over time
- *"5-year risk of event for Patient X is 0.57"*

$$P(T \leq t) = \int_0^t P(T > u-)\alpha(u)du$$

Motivation

What should we model in a survival analysis?

Interested in risk factors?

Model cause-specific hazard

- Quantify effects of covariates on the rate of the outcome in subjects

Interested in predicting patient outcome?

Model cause-specific cumulative incidence

- Estimate patient prognosis
- Inform clinical patient management

Competing Risks Models: Cox Proportional Hazards

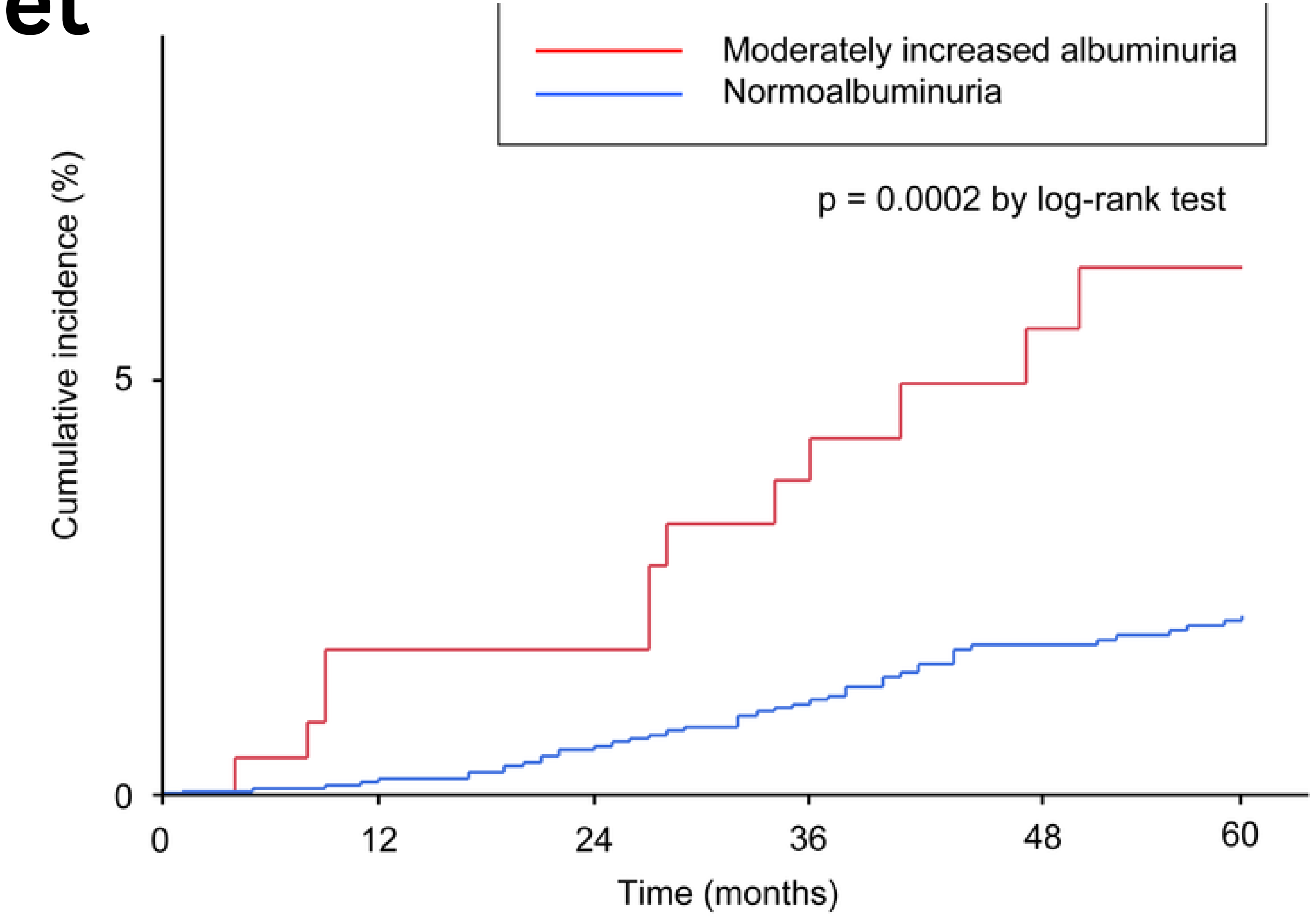
$$h(t|x_i) = h_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta})$$

- Models **cause-specific hazards:** subjects who are event free
- **Flexible semi-parametric approach:** measures relative risk through hazards ratio
- **Proportional Hazards Assumption:** Baseline hazard not explicitly modelled

$$\frac{h(t|X_1)}{h(t|X_2)} = \frac{h_0(t)\cdot\exp(\beta_1 X_1)}{h_0(t)\cdot\exp(\beta_2 X_2)} = \frac{\exp(\beta_1 X_1)}{\exp(\beta_2 X_2)}$$

Competing Risks Models: Cox Proportional Hazards

- Treats competing risk as censored
- **To estimate CIF:** Have to estimate baseline hazard separately
- Produces **non-parametric,** step-wise estimates of cumulative incidence: **difficult to interpret**



Number at risk

Moderately increased albuminuria	237	226	209	189	143	128
Normoalbuminuria	2824	2704	2516	2154	1779	1618

Competing Risks Models: CIF Models

- E.g. Fine Gray model
- **Model CIF through sub-distribution hazards:** Consider **event-free** and individuals who have **experienced competing event**
- **One-to-one relationship with CIF**
- Produces **non-parametric** CIF estimates that account for competing risks
- Also produces step-wise estimates of CIF

Overview of competing risks models

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Treat competing risks as censored

CIF Models

☑ Quantify clinical prognosis

☑ Account for competing risks

☑ Produce step-wise estimates of CIF (difficult to interpret)

Overview of competing risks models

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Treat competing risks as censored

CIF Models

☑ Quantify clinical prognosis

☑ Account for competing risks

☑ Produce step-wise estimates of CIF
(difficult to interpret)

Overview of competing risks models

Cause-Specific Hazards Models

- ☑ Quantify Risk Factors: easy to interpret
- ☑ Treat competing risks as censored

CIF Models

- ☑ Quantify clinical prognosis
- ☑ Account for competing risks

☑ Produce step-wise estimates of CIF
(difficult to interpret)

Overview of competing risks models

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Treat competing risks as censored

CIF Models

☑ Quantify clinical prognosis

☑ Account for competing risks

☑ Produce estimates of CIF that are smooth in time: easy to interpret

Proposed Method

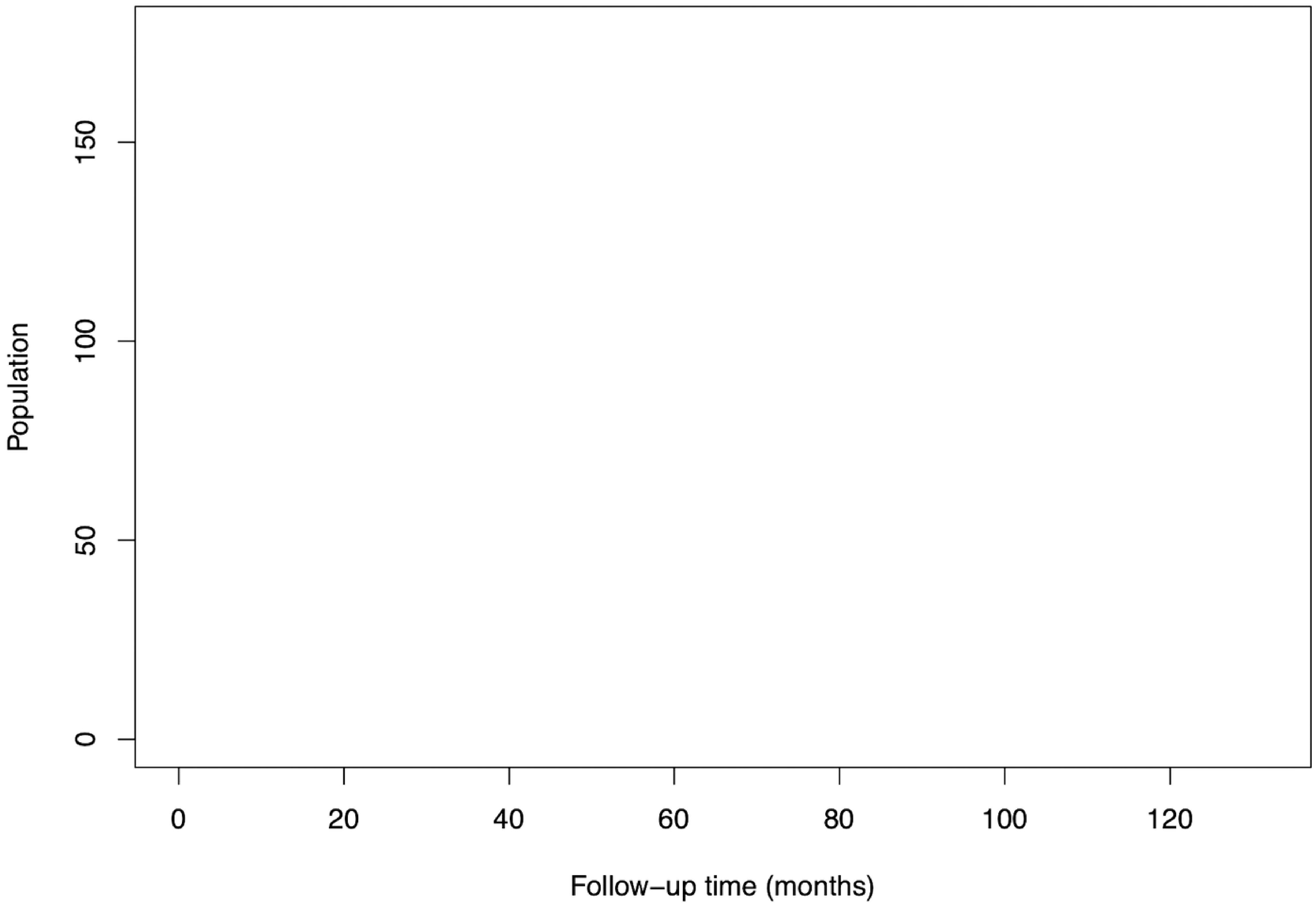
Solution: Casebase framework (Bhatnagar et al., 2020)

- Models the cause-specific hazard directly using (smooth) parametric distribution families

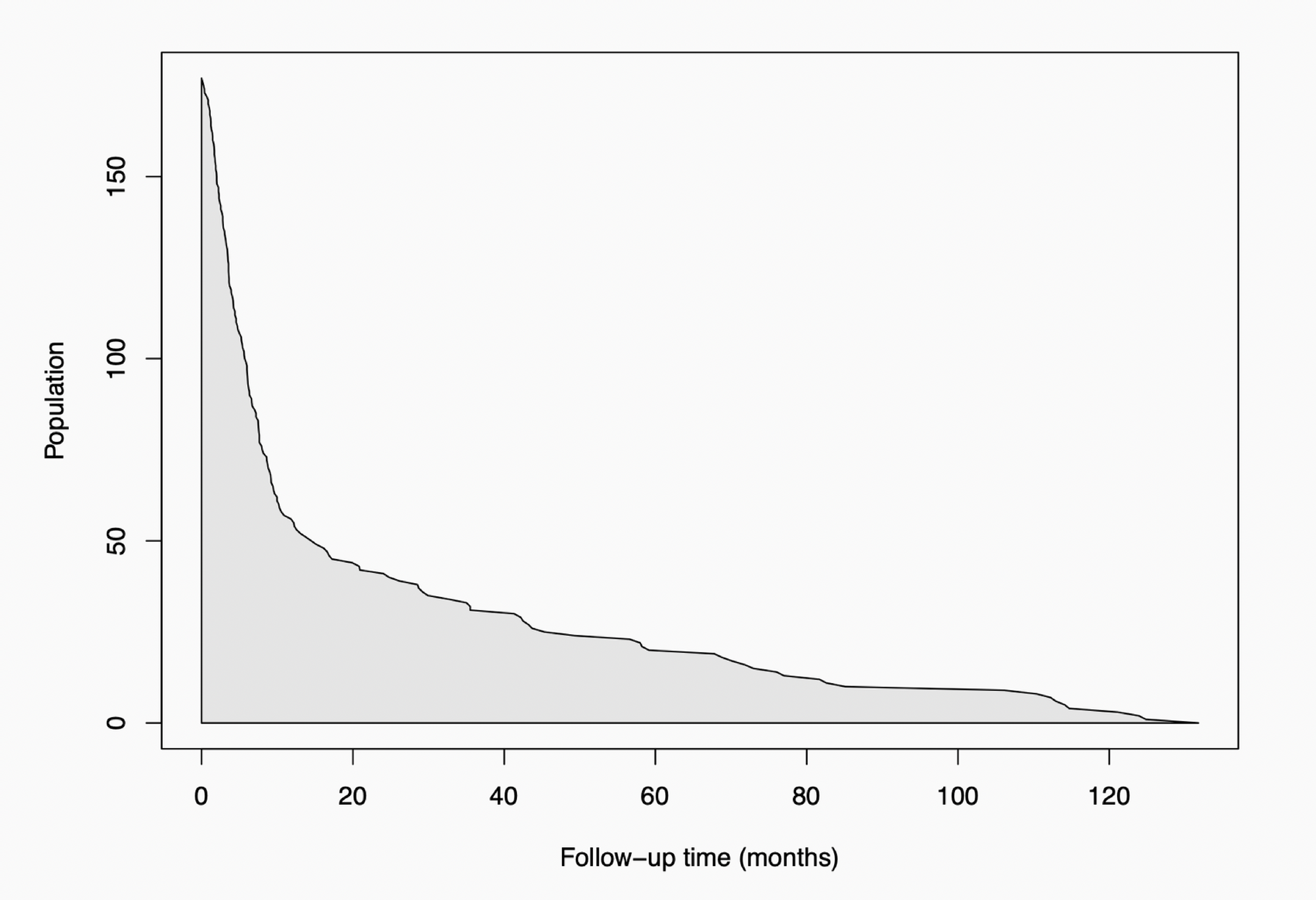
- Produces smooth CIF curves, adjusting for competing risks

- Based on Hanley & Miettinen’s (2009) case base sampling method

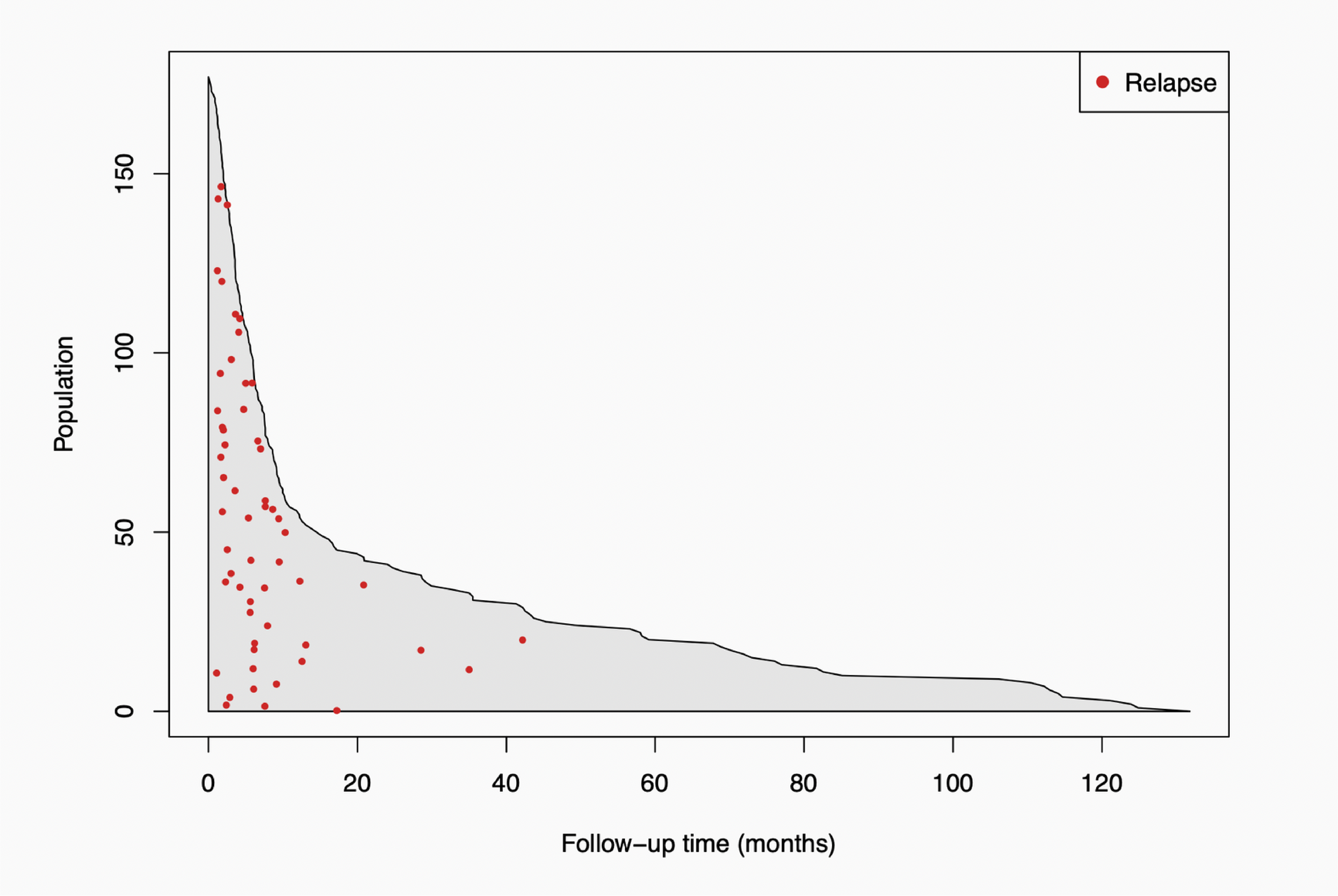
Casebase sampling



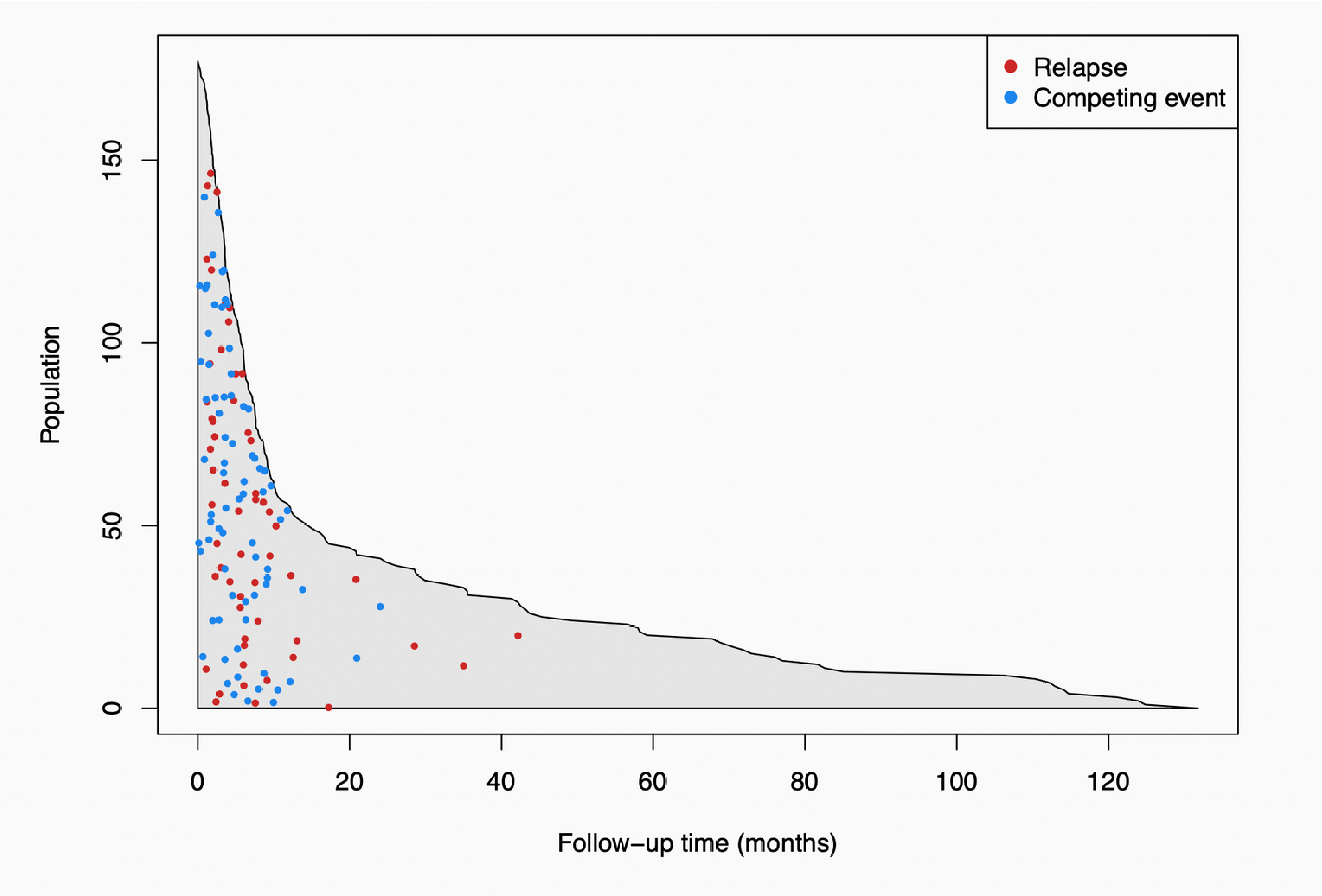
Casebase sampling



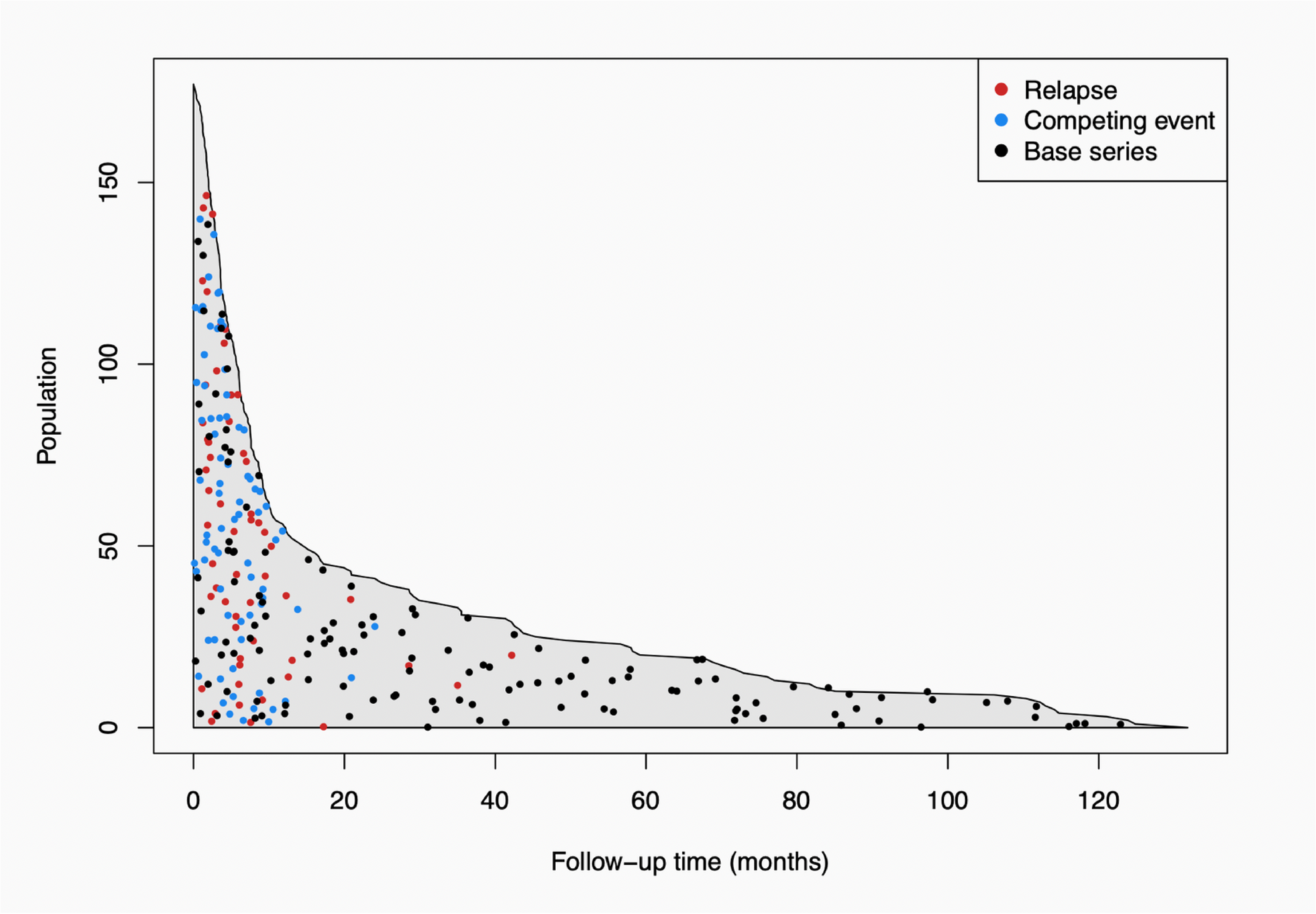
Casebase sampling



Casebase sampling



Casebase sampling



Casebase framework

Let $N_i(t) \in \{0, 1, 2\}$ be counting processes corresponding to the event for individuals $i = 1, \ldots n$

We model the hazard function to satisfy:

$$\lambda_i(t)dt = E[dN_i(t) \mid \text{past}].$$

Casebase framework

We model the hazard function to satisfy:

$$\lambda_i(t)dt = E[dN_i(t) \mid \text{past}]$$

If the hazard function $\lambda_i(t;\theta)$ is parameterized in terms of θ we can define an estimator by maximizing the likelihood:

$$L_0(\theta) = \prod_{i=1}^{n} \exp\left\{-\int_0^{\min(t_i,\tau)} \lambda_i(t;\theta)dt\right\} \prod_{i=1}^{n} \prod_{t\in[0,\tau)} \lambda_i(t;\theta)^{dN_i(t)},$$

Casebase framework

- By conditioning on person-moments through case-base sampling, we can avoid computing the integral

We can define an estimating equation for θ as follows:

$$L(\theta) = \prod_{i=1}^{n} \prod_{t \in [0,\tau]} \left(\frac{\lambda_j(t)^{dN_j(t)}}{\rho_i(t) + \sum_{j=1}^{J} \lambda_j(t)} \right)$$

where $$\rho_i(t) = \frac{\text{N in base series}}{\text{Total population time of study-base}}$$

- Corresponds to multinomial likelihood with offset $\log(1/\rho_i(t))$.

Casebase Estimation: Multinomial Regression

- Multinomial Regression parameterization:

$$\log \frac{\Pr(G=l|x,t)}{\Pr(G=K|x,t)} = \beta_{0l} + x^T\beta_l + \log(B/b), \quad l = 1, \ldots\ldots, K-1$$

- **glmnet** uses symmetric parameterization: does not estimate offset
- Optimize using stochastic variance reduced gradient (SVRG) (Johnson and Zhang., 2013): fast convergence for *p* > *n*
- Implemented in **mtool** package

$$\min_{\theta \in \mathbb{R}^p} -\ell(\theta) + \sum_{i=1}^{p} w_j \lambda \left(\frac{1-\alpha}{2} \sum_{k=1}^{p} |\theta_{j_k}| + \frac{\alpha}{2} \sum_{k=1}^{p} |\theta_{j_k}|^2 \right)$$

Simulation Study: Variable Selection

Data Generation: Survival Data

Competing Risks Survival Data

- Survival times generated from two exponential distributions using inverse transform sampling (one-year time period)

$$t_i = \frac{-\log(u_1)}{0.1 \cdot \exp(X\beta)}, i = 1, 2$$

- Cause-indicator **(1 - Cause of interest, 2 - Competing Risk)** generated from binomial experiment

$$\text{Binomial}[n, (1-p)^{\exp(X\beta_1)}]$$

where $p = 0.5$

- Uniform censoring: $U[0, M]$

- ~ **44 %** - Cause of interest, ~ **42 %** - Competing Risk, ~ **15 %** Censoring Rate

Data Generation: Covariate Generation

- True predictors generated from MVN with $\boldsymbol{\mu} = 0$ and pairwise correlations $\rho = 0.5$
- Noise predictors generated from MVN with $\boldsymbol{\mu} = 0$ and pairwise correlations $\rho = 0.1$
- We set $\beta_1 = (0, 1)^T$ and $\beta_2 = -\beta_1$
- N = 400
- Ratio of number of predictors (p)/ True predictors: 120/50, 1000/50

Data Generation: Simulation Settings

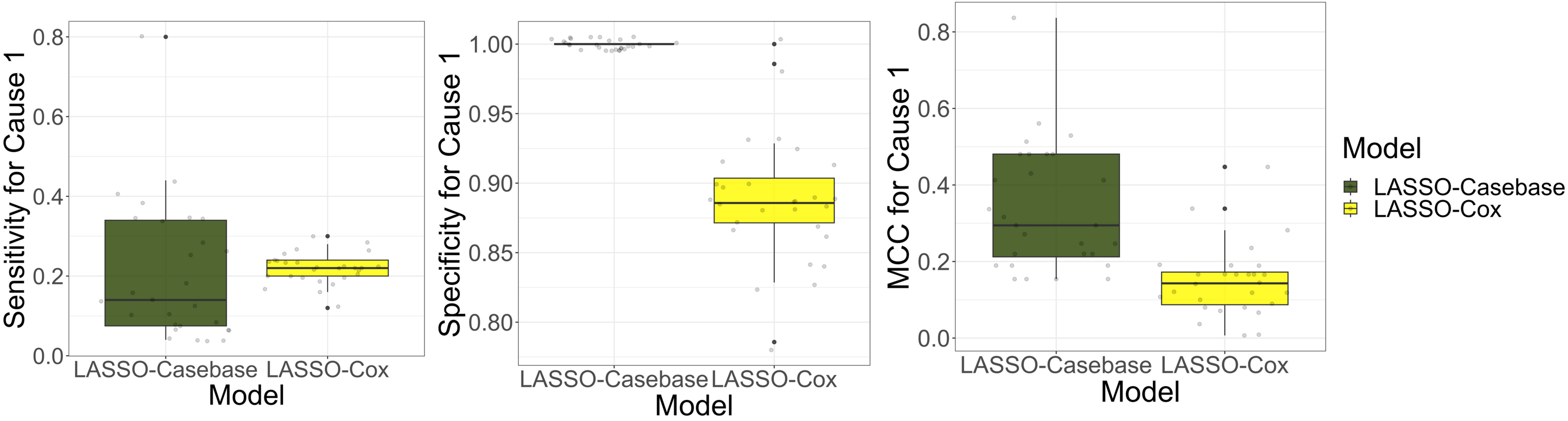
Compare Cause-Specific Hazard Models: Look at Cause 1

1. **Pen. Case-base with LASSO penalty (LASSO-Casebase)**

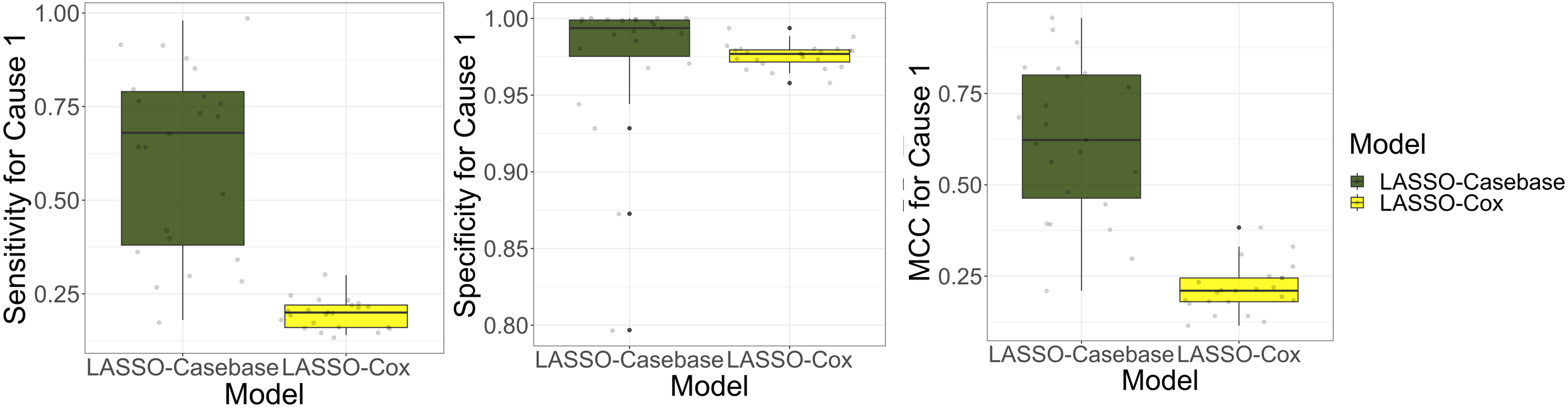
2. **Pen. cox with LASSO penalty (LASSO-Cox)**

- Tune Case-base using 5-fold cross-validation and Cox using 10-fold cross-validation, select lambda.min
- Time variable is transformed into log(Time) to model Weibull hazard in case-base and is not penalized
- **Comparison Metrics:** Sensitivity, Specificity, Matthew's Correlation Coefficient (MCC) (FP = FN = 0: +1, TP = TN = 0: -1)

Results: N = 400, p = 120, Tp = 50



Results: N = 400, p = 1000, Tp = 50



Simulation Study: CIF Prediction

Data Generation: Simulation Settings

Compare Cause-Specific and CIF Models:

- X generated from IID MVN

1. Pen. Case-base with LASSO penalty (LASSO-Casebase)
2. Pen. cox with LASSO penalty (LASSO-Cox)
3. Boosted Fine-Gray Model (FineBoost)

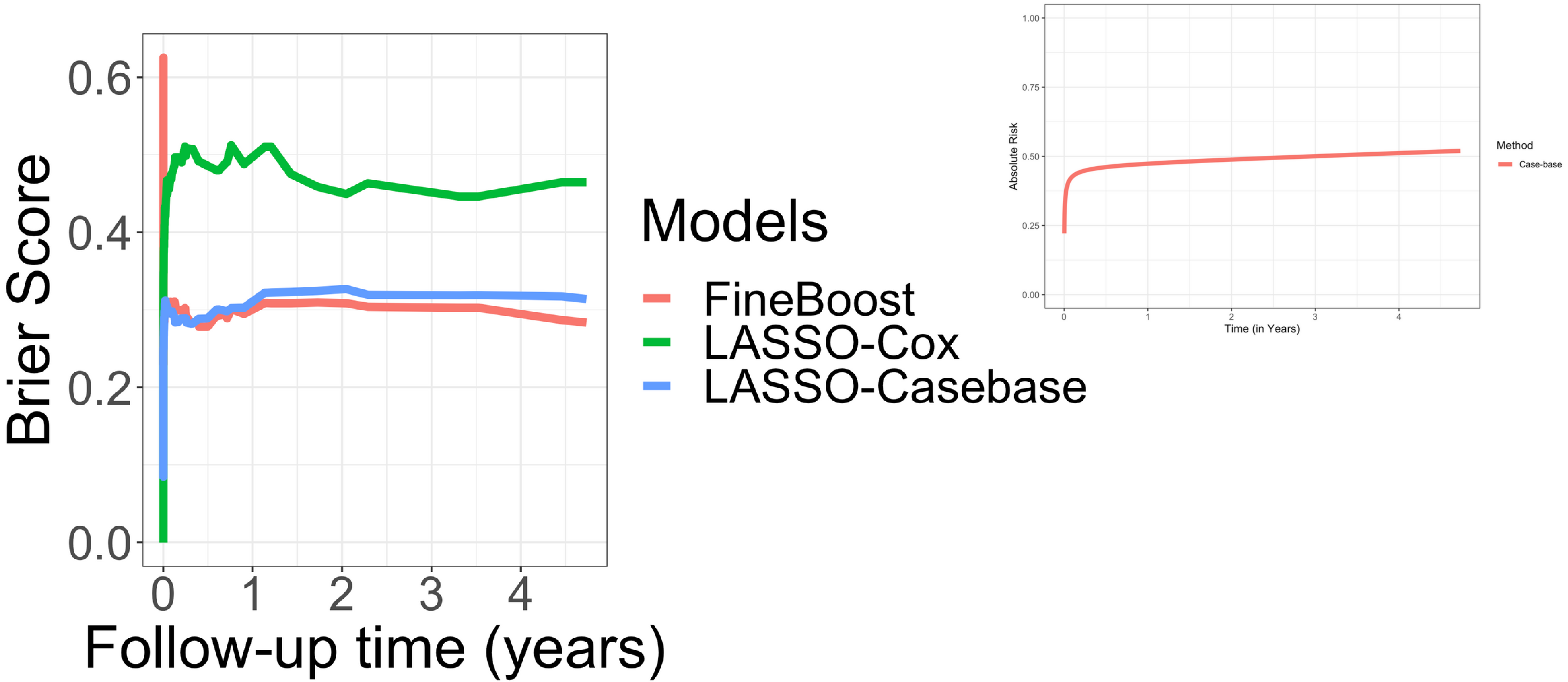
- **Comparison Metrics:** Time-dependent Brier Score

$$B(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\widetilde{S}_i(t)}\sum_{j=1}^{m} I(Y_i(t)=j)\Big(I(Y_i(t)=j) - \widehat{P}(Y_i(t)=j)\Big)^2$$

δ_i : event indicator variable

$\widetilde{S}_i(t)$:Estimated censoring survival function for individual *i* at time t (KM)

Results: N = 400, p = 120, Tp = 50



Conclusions and Next Steps

Overview of competing risks models

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Treat competing risks as censored

CIF Models

☑ Quantify clinical prognosis

☑ Account for competing risks

☑ Produce step-wise estimates of CIF (difficult to interpret)

Overview of competing risks models

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Treat competing risks as censored

CIF Models

☑ Quantify clinical prognosis

☑ Account for competing risks

☑ Produce step-wise estimates of CIF (difficult to interpret)

Overview of competing risks models

Cause-Specific Hazards Models	**CIF Models**
☑ Quantify Risk Factors: easy to interpret	☑ Quantify clinical prognosis
☑ Treat competing risks as censored	☑ Account for competing risks

☑ Produce estimates of CIF that are smooth in time: easy to interpret

Conclusion

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Performs comparably to cox in n > p and outperforms cox in p >n

CIF Models

☑ Quantify clinical prognosis

☑ Account for competing risks

☑ Produce step-wise estimates of CIF (difficult to interpret)

Overview of competing risks models

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Treat competing risks as censored

CIF Models

☑ Performs comparably to CIF models in prediction

☑ Produce step-wise estimates of CIF (difficult to interpret)

Overview of competing risks models

Cause-Specific Hazards Models

☑ Quantify Risk Factors: easy to interpret

☑ Treat competing risks as censored

CIF Models

☑ Quantify clinical prognosis

☑ Account for competing risks

☑ Produce estimates of CIF that are smooth in time: easy to interpret

Next Steps

- Deal with uniform censoring randomness
- Bootstrapped confidence intervals for the Brier score (using .632+ rule)
- More CIF comparison models (Direct Binomial)
- **If time:** analysis on dataset (p > n) 2000 genes, 400 observations and time-event data on Bladder Cancer

Current Struggles:

- p > n cases running out of memory on Compute Canada :(